Reading Comprehension For Multimedia Question Answering

M2 Internship

Keywords: visual question answering, natural language processing, computer vision, machine learning, artificial intelligence.

Context

The internship takes place in the framework of the MEERQAT project[1], which focuses on Multimedia Question Answering (MQA). This task consists in answering questions grounded in a visual context. For instance, while watching a film, one can wonder "*In which movie did I already see this actress?*" or "*How many Oscar did she won?*". It is related to Visual Question Answering (VQA, Antol et al., 2015). However, VQA questions relate to the content of the image, such as the color of an object or the number of objects (e.g. one could ask "*What color is his shirt?*" from the first image in Table 1), while MQA focuses on finding answers in text, but with the help of images associated with the questions (see Table 1 for some examples).

Research problem

Question Answering is usually split into two steps: Information Retrieval for selecting a restricted set of documents or passages from a large collection of documents and Reading Comprehension for extracting answers to the questions in the retrieved documents. The internship will focus on the second step, relying on the work already done in the MEERQAT project for the multimedia retrieval of documents. For a few years, Reading Comprehension has been addressed with attention-based neural networks, which take as input the question and a candidate passage of text where one might find the answer (Chen et al., 2017). This approach became ubiquitous since the dawn of pre-trained language models such as BERT (Devlin et al., 2019).

However, in the case of MQA, a text-only strategy for Reading Comprehension is not always sufficient for choosing between the answers extracted from two candidate passages. For instance, in the case of Table 1's question (b), another candidate passage could be *"Príncipe de Asturias was a steam ocean liner, built in Scotland for the Spanish Naviera Pinillos. She was launched in 1914 and wrecked in 1916 with the loss of at least 445 lives."* and would lead to giving *1914* as an answer instead of *1968*.

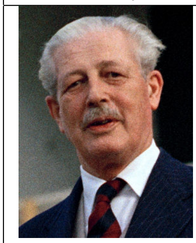
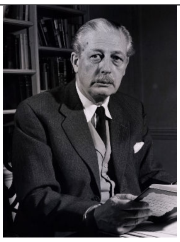
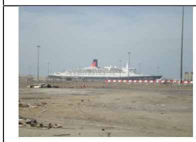
Query	**Knowledge Base**
	
(a) "Which constituency did this man represent when he was Prime Minister?"	"Macmillan indeed lost Stockton in the landslide Labour victory of 1945, but returned to Parliament in the November 1945 by-election in **Bromley**."
	
(b) "In which year did this ocean liner make her maiden voyage?"	"Queen Elizabeth 2, often referred to simply as QE2, is a floating hotel and retired ocean liner built for the Cunard Line which was operated by Cunard as both a transatlantic liner and a cruise ship from **1969** to 2008."

Table 1: Example of questions along with their contextual image and answer source.

Objectives

The main objective of the internship is to define, implement, and evaluate methods, in the context of MQA, for taking into account the information brought by images in the Reading Comprehension task.

Two main research directions will be considered in this context:

[1] www.meerqat.fr

- a late fusion approach relying on the results of the multimedia Information Retrieval step to rerank candidate answers with respect to images;
- a more early fusion approach integrating images in the reader to allow contextual disambiguation.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile. IEEE.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Internship conditions

The internship will be supervised by Paul Lerner along with Olivier Ferret and Camille Guinaudeau and will take place at LISN[2]. LISN is an interdisciplinary laboratory resulting from the merge of LIMSI and LRI in 2021. It is associated with CNRS and Université Paris-Saclay and includes 16 research teams and 380 people. The intern will be located at *bât 507, Rue du Belvedère, F-91405 Orsay cedex*.

- Remuneration: around 600€ along with the refund of half the Navigo (public transport) card.
- Starting date: the internship is expected to start from March 2022 but could begin earlier.
- Duration: 5-6 months.

Requirements

We are looking for an M2 student in Natural Language Processing, Computer Vision or Machine Learning. The intern is expected to be proficient in programming, especially in the Python language, and to have already worked under Linux. They should also have experience with a deep learning framework, preferably PyTorch.

Application

Please send a resume along with a cover letter (in French or English) and grade transcripts for the last two years to Paul Lerner at paul.lerner@lisn.upsaclay.fr. Examples of projects (e.g. via GitHub) is a plus.

[2] www.lisn.upsaclay.fr